

Poor Hygiene Corrupts Analysis

*By Jim Wheaton (co-author)
Principal, Wheaton Group*

Original Version of an article that appeared in the June 2, 2003 issue of “DM News”

Some time ago, a data audit was performed for a direct marketing company. It soon became apparent that the direct marketer’s order entry procedures were extremely lax. Many records were found with outright address element misspellings and omissions. Others displayed technically correct but problematic phenomena, such as the input of initials rather than full first names. Consider, for example, the following three pairs of New York City records:

- Ben Rosen and Ben Orsen at 1407 Madison Avenue.
- R. Happle and Robert Appel at 130 East End Avenue.
- Ms. K. Mahoney and Katherine Maloney at 829 Park Avenue.

The individuals represented by these records reside in multiple family dwelling units. Nevertheless, the records contain no apartment number information.

Do each of these record pairs represent the same individual? And, how do hygiene issues affect response analysis and data mining? Both questions will be explored in this article.

Definition of a Duplicate

Consider the following two pairs of records:

Record 1	Record 2
James Wheaton 151 Thurton Drive New Canaan, CT 06840	James Wheaton 151 Thurton Drive New Canaan, CT 06840

Record 1	Record 2
Beth Wilson 52 Devils Garden Road New Canaan, CT 06854	Buffy Walters 52 Devils Garden Road New Canaan, CT 06854

It appears obvious that Pair #1 contains the duplicate record, and that Pair #2 represents two different individuals at the same address. Surprisingly, however, Pair #2 contains the duplicate.

Pair #1 consists of two different people, a father and his son, with their respective suffixes (Jr. and III) deleted. It was a constant source of confusion for me while growing up, as is the case with any son who is named after his father.

In Pair #2, the first record represents a married woman's professional name, comprised of her given name and maiden surname. As with many women, Beth did not change her name professionally when she married. She decided to retain the name with which all of her co-workers and associates were familiar. The second record contains Beth's nickname and married surname. In her personal life, she opted for her surname to correspond with her husband's.

These two pairs of records illustrate that there is no way to be 100 percent correct when it comes to defining duplicates. Therefore, it is best to incorporate specialized hygiene technology into operational systems, and improve order entry procedures, in order to minimize the occurrence of potential duplicate situations. In this way, back-end cleanup is employed only as a last resort. Several appropriate hygiene technologies and order entry procedures will be discussed in the final section of this article.

Ramifications of Poor Hygiene

Consider what happens whenever two legitimate duplicates are not consolidated. When doing matchback response analysis, each unconsolidated duplicate represents one less attributed order. The higher the number of unattributed orders, the more difficult it is to accurately quantify the performance of lists and list segments.

For customers, an actual multi-buyer will appear to be two separate – and less desirable – single-buyers, which will reduce the effectiveness of any statistics-based predictive model. After all, “pseudo” single-buyers will purchase more frequently in the future than expected – just as “pseudo” multi-buyers will purchase less frequently. Likewise for lifetime value analysis, one relatively valuable customer will appear to be two not-so-valuable individuals. And, the opposite effect will occur whenever an inappropriate record consolidation takes place.

Even with records that are likely to be duplicates, sloppy order entry procedures cause problems. Consider, for example, the earlier “Ben Rosen” and “Ben Orsen” record pair. Which is the correct surname for this individual? This is important because people appreciate being correctly referred to. When, they are not, the logical result is a lowered likelihood to place future orders.

Superior Record Hygiene: Even More Important Than 20 Years Ago

Twenty years ago, most direct marketing orders for many companies arrived via the mail. In this long-ago world, the majority of orders could be directly attributed to a promoted name and address.

This is because most people filled out the order form that accompanied the direct mail piece. These order forms, in turn, contained the proper name and address of the prospect or customer. Of the minority of orders that were not directly attributable, most were the result of “passalong mail,” where a friend or relative was inspired by the direct marketing piece to place an order.

Today, almost all orders are handled by inbound call centers and e-commerce sites. Unfortunately, when it comes to name and address entry, many call centers display less-than-rigorous standards. Glaring misspellings are common, as are omissions of address elements. Likewise, the capture of valid key codes often is not much better. And, e-commerce orders present their own name and address quality challenges. Frequently, for example, no more than 20 to 25% contain a key code.

This attenuated linkage between those who are promoted and those who respond makes it difficult to analyze direct marketing campaigns. Barriers to analysis, in turn, increase the probability that incorrect rollout decisions will be made, and that predictive models and lifetime value analysis will not reach their full potential.

Maximizing the Quality of Response Data

Fortunately, techniques exist to maximize the quality of response data. Unique ID’s can be applied to promoted records. When applied to prospects, these are often referred to as “finder numbers.” Call center reps can request this unique ID, which acts as a “hard” link between the order and the promoted record. Likewise, the ID can be requested during e-commerce sessions.

Ideally, when the ID is input by the call center rep into the operational system, the name and address of the customer or prospect will appear on the screen. Similarly, the same steps can occur during e-commerce sessions. At this time, the customer can be queried about any changes, including whether he or she is a “passalong” order. If this is the case, then a “hard link” has been established between the response vehicle and “indirectly promoted” new customer, which is very helpful in subsequent response analysis and data mining.

For direct marketers whose size is sufficiently large to justify the investment, real-time hygiene technologies can also be integrated into their operational infrastructure. First, customer and prospect contact lists are loaded into the operational system. If incoming orders have different address data, technology can be used to “screen on the fly” for address elements that do not meet USPS standards. For example, out-of-range street numerics can be flagged for follow-up investigation by the call center rep or through an e-commerce screen. Such technologies significantly reduce the volume of problematic name and address information that is input to operational systems.

Unfortunately, many small to mid-sized direct marketers cannot justify major investments in up-front hygiene technology. However, name and address input quality can be significantly enhanced by the incorporation of simple but cost-effective order entry procedures. For example, rigorous name and address input standards can be established for call center reps, and the performance of individual reps tracked by the back-end matching of orders against prospect and customer lists. Those reps who consistently achieve high standards become eligible for performance bonuses.

Jim Wheaton is a Principal at Wheaton Group, which specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. He is also co-founder of Data University. Jim can be reached at 919-969-8859, or jim.wheaton@wheatongroup.com.